Tanta University

Department: **Computers and Control Engineering**
(20 marks)

Faculty of Engineering

Course Title: **Information Systems Design** Course Code: CCE4235 4th year
Date: 5.4.2017 (Second term) Allowed time: 1 hrs

Answer the following questions:

Question No. 1 **(10 marks)**

1. Consider the following Training Data Set:
   Using Naïve Bayesian Classifier, based on the object's attributes Red Domestic SUV label this object as stolen or not. **(5 marks)**

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2. What are the key skill sets and behavioral characteristics of a Data Scientist? **(2 marks)**
3. Discuss the phases of the Data Analytics Lifecycle in the context of the mini case: Churn Prediction for Retail Banking. **(3 marks)**

Question No. 2 **(10 marks)**

1. A psychologist was interested in whether different TV shows lead to a more positive outlook on life. People were split into 4 groups and then taken to a room to view a program. The four groups saw: The Muppet Show, Futurama, The News, No program. After the program a blood sample was taken and serotonin levels measured (remember more serotonin means happier). The levels are given below for the four different groups. Carry out a one-way ANOVA to test the hypothesis that the treatments will have different effects. **(5 marks)**

	The Muppet Show	Futurama	BBC News	No Program
	11	4	4	7
	7	8	3	7
	8	6	2	5
	14	11	2	4
	11	9	3	3
	10	8	6	4
	5			4
				4
Mean	9.43	7.67	3.33	4.75
Variance	8.95	5.87	2.27	2.21
Grand Mean			6.30	
Grand Variance			10.06	

2. A database consisting of 9 transactions containing five items is shown in the table below.

a) Apply Apriori algorithm (let the minimum support= 22%) to find all the frequent item sets in the database. **(2 marks)**

b) Use these frequent item sets and the minimum confidence constraint (let the minimum confidence= 70%) to form the association rules. **(3 marks)**

TID	List of items
T_1	I_1,I_2,I_5
T_2	I_2,I_4
T_3	I_2,I_3
T_4	I_1,I_2,I_4
T_5	I_1,I_3
T_6	I_2,I_3
T_7	I_1,I_3
T_8	I_1,I_2,I_3,I_5
T_9	I_1,I_2,I_3

Best wishes

Dr. Sherin El Gokhy

Question No. 1

1 Naïve Bayesian

- Classify the object as Stolen or not based on object attributes Red, Domestic, SUV → A

$P(\text{Stolen}) = 0.5 \qquad P(\text{not Stolen}) = 0.5$

$$P(\text{Red} \mid \text{Stolen}) = \frac{P(\text{Red} \cap \text{Stolen})}{P(\text{stolen})} = \frac{0.3}{0.5} = 0.6$$

$$P(\text{Domestic} \mid \text{Stolen}) = \frac{P(\text{Domestic} \cap \text{stolen})}{P(\text{Stolen})} = \frac{0.2}{0.5} = 0.4$$

$$P(\text{SUV} \mid \text{Stolen}) = \frac{P(\text{SUV} \cap \text{Stolen})}{P(\text{Stolen})} = \frac{0.1}{0.5} = 0.2$$

$$P(\text{Red} \mid \text{not}) = \frac{P(\text{Red} \cap \text{not})}{P(\text{not})} = \frac{0.2}{0.5} = 0.4$$

$$P(\text{Domestic} \mid \text{not}) = \frac{P(\text{Domestic} \cap \text{not})}{P(\text{not})} = \frac{0.3}{0.5} = 0.6$$

$$P(\text{SUV} \mid \text{not}) = \frac{P(\text{SUV} \cap \text{not})}{P(\text{not})} = \frac{0.3}{0.5} = 0.6$$

$$P(\text{Stolen} \mid A) = \frac{P(A \mid \text{stolen})\, P(\text{stolen})}{P(A)}$$

$$P(\text{stolen} \mid A) \propto P(\text{Red} \mid \text{stolen})\, P(\text{Dome.} \mid \text{stolen})\, P(\text{SUV} \mid \text{stolen})\, P(\text{stolen})$$

$$P(\text{stolen} \mid A) \propto 0.6 * 0.4 * 0.2 * 0.5 = \boxed{0.024}$$

$$P(\text{not} \mid A) = \frac{P(A \mid \text{not})\, P(\text{not})}{P(A)}$$

$$P(\text{not} \mid A) \propto P(\text{Red} \mid \text{not})\, P(\text{Dome.} \mid \text{not})\, P(\text{SUV} \mid \text{not})\, P(\text{not})$$

$$P(\text{not} \mid A) \propto 0.4 * 0.6 * 0.6 * 0.5 = \boxed{0.072}$$

The object class is not stolen because $P(\text{not} \mid A) > P(\text{stolen} \mid A)$

Question No. 2

1 ANOVA

→ Step 1 : The Mean

$$m_1 = \frac{11+7+8+14+11+10+5}{7} = 9.43$$

$$m_2 = \frac{4+8+6+11+9+8}{6} = 7.67$$

$$m_3 = \frac{4+3+2+2+3+6}{6} = 3.33$$

$$m_4 = \frac{7+7+5+4+3+4+4+4}{8} = 4.75$$

$$m_0 = \frac{m_1+m_2+m_3+m_4}{4} = 6.30$$

→ Step 2 : Sum of squar

$$SS_{\text{with-in}} = \sum_j (x_1^j - m_1)^2 + \sum_j (x_2^j - m_2)^2 + \sum_j (x_3^j - m_3)^2 + \sum_j (x_4^j - m_4)^2$$

$$= 53.7143 + 29.3334 + 11.33 + 15.50$$

$$SS_{\text{with-in}} = 109.88$$

$$SS_{\text{total}} = \sum_i \sum_j (x_i^j - m_0)$$

$$SS_{\text{total}} = 261.63$$

$$SS_{\text{Between}} = SS_{\text{total}} - SS_{\text{between}} = 151.75$$

→ Step 3 : F

K → number of attributes

N → number of items

$$S_B^2 = \frac{SS_{\text{Between}}}{K-1} = \frac{151.75}{4-1} = 50.583$$

$$S_w^2 = \frac{SS_{\text{within}}}{N-K} = \frac{109.88}{27-4} = 4.777$$

$$F = \frac{S_B^2}{S_w^2} = \frac{50.583}{4.777} = \boxed{10.589}$$

$F > 1$

reject Null hypo.

Other Solution

→ using data in the table

$$S_B^2 = \frac{1}{K-1} \sum n_i \, (m_i - m_o)^2$$

$$= \frac{1}{4-1}\Big[7\,(9.43-6.3)^2 + 6\,(7.67-6.3)^2 + 6\,(3.33-6.3)^2 + 8\,(4.75-6.3)^2 \Big]$$

$$= \frac{151.9851}{4-1}$$

$$\boxed{S_B^2 = 50.6617}$$

V_i → Variance

$$S_W^2 = \frac{1}{N-K} \sum n_i * V_i$$

$$= \frac{1}{27-4}\left[7*8.95 + 6*5.87 + 6*2.27 + 8*2.21 \right]$$

$$= \frac{129.17}{27-4}$$

$$\boxed{S_W^2 = 5.616}$$

$$F = \frac{S_B^2}{S_W^2} = \boxed{9.0209}$$

$F > 1$

Reject Null hypo.

ملاحظة

- إختلاف قيمة S_W^2 بسبب أن الـ Variance المحسوبة في الجدول قيمة ليست صحيحة

Question No. 2

(2) Apriori Algorithm

- min Support 22 %
- min Confidence 70 %

Step 1:

Frequent Items	Count	Support
I1	6	66.67 %
I2	7	77.78 %
I3	6	66.67 %
I4	2	22.22 %
I5	2	22.22 %

support = Count / no. of Transactio.

no. of Transaction = 9

→ We should Prune Items with Support < 22%

→ There is no Item to be pruned

Step 2: Item pairs

Frequent Items	Count	Support	
I1, I2	4	44.44 %	
I1, I3	4	44.44 %	
~~I1, I4~~	1	11.11 %	→ Prune
I1, I5	2	22.22 %	
I2, I3	4	44.44 %	
I2, I4	2	22.22 %	
I2, I5	2	22.22 %	
~~I3, I4~~	0	0 %	→ Prune
~~I3, I5~~	1	11.11 %	→ Prune
~~I4, I5~~	0	0 %	→ Prune

Step 3 :

→ make Frequent items from 3 item → The order is not important

→ Consider if Support (X, y) < 22 %
then Support (X, y, z) will be Less than 22%
So We will not take X, y, z

Frequent items	Count	Support	We Ignore
I1, I2, I3	2	22.22 %	I1, I2, I4
I1, I2, I5	2	22.22 %	I1, I3, I4
			I1, I3, I5
			I2, I3, I4
			I2, I3, I5

Step 4 :

Frequent items	Count	Support	
~~I1, I2, I3, I5~~	1	11.11 %	→ Prune

* We have run out of support

→ The Algorithm will Stop after step 4

Finally : Rules Confidence → we can find count in tables of the previous steps

Rule	Set → cnt	Set → cnt	Confidence
I1 → I2	I1 6	I1, I2 4	4/6 = 67%
I2 → I1	I2 7	I1, I2 4	4/7 = 57%
I1 → I3	I1 6	I1, I3 4	4/6 = 67%
I3 → I1	I3 6	I1, I3 4	4/6 = 67%
I1 → I5	I1 6	I1, I5 2	2/6 = 33%
I5 → I1	I5 2	I1, I5 2	2/2 = 100%

Rule	Set	cnt	Set	cnt	Confidence
$I_2 \rightarrow I_3$	I_2	7	I_2, I_3	4	4/7 = 57%
$I_3 \rightarrow I_2$	I_3	6	I_2, I_3	4	4/6 = 67%
$I_2 \rightarrow I_4$	I_2	7	I_2, I_4	2	2/7 = 29%
$I_4 \rightarrow I_2$	I_4	2	I_2, I_4	2	2/2 = 100%
$I_2 \rightarrow I_5$	I_2	7	I_2, I_5	2	2/7 = 29%
$I_5 \rightarrow I_2$	I_5	2	I_2, I_5	2	2/2 = 100%
$I_1 \rightarrow I_2, I_3$	I_1	6	I_1, I_2, I_3	2	2/6 = 33%
$I_2, I_3 \rightarrow I_1$	I_2, I_3	4	I_1, I_2, I_3	2	2/4 = 50%
$I_2 \rightarrow I_1, I_3$	I_2	7	I_1, I_2, I_3	2	2/7 = 29%
$I_1, I_3 \rightarrow I_2$	I_1, I_3	4	I_1, I_2, I_3	2	2/4 = 50%
$I_3 \rightarrow I_1, I_2$	I_3	6	I_1, I_2, I_3	2	2/6 = 33%
$I_1, I_2 \rightarrow I_3$	I_1, I_2	4	I_1, I_2, I_3	2	2/4 = 50%
$I_1 \rightarrow I_2, I_5$	I_1	6	I_1, I_2, I_5	2	2/6 = 33%
$I_2, I_5 \rightarrow I_1$	I_2, I_5	2	I_1, I_2, I_5	2	2/2 = 100%
$I_2 \rightarrow I_1, I_5$	I_2	7	I_1, I_2, I_5	2	2/7 = 29%
$I_1, I_5 \rightarrow I_2$	I_1, I_5	2	I_1, I_2, I_5	2	2/2 = 100%
$I_5 \rightarrow I_1, I_2$	I_5	2	I_1, I_2, I_5	2	2/2 = 100%
$I_1, I_2 \rightarrow I_5$	I_1, I_2	4	I_1, I_2, I_5	2	2/4 = 50%

The Rules that we have

1) If I5 Then I1
2) if I4 Then I2
3) if I5 Then I2
4) if I2, I5 Then I1
5) if I1, I5 Then I2
6) if I5 Then I1, I2